(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 November 2002 (14.11.2002)

PCT

(10) International Publication Number
WO 02/090569 A2

(51) International Patent Classification⁷: C12Q

(21) International Application Number: PCT/US02/14436

(22) International Filing Date: 7 May 2002 (07.05.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/289,068 7 May 2001 (07.05.2001) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/289,068 (CIP)
Filed on 7 May 2001 (07.05.2001)

(71) Applicant (*for all designated States except US*): **CURAGEN CORPORATION** [US/US]; 555 Long Wharf Drive, 11th floor, New Haven, CT 06511 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): **BADER, Joel, S.** [US/US]; 1127 High Ridge Road, #107, Stamford, CT 06905 (US). **SHAM, Pak** [GB/US]; 46 Larkhall Rise, Clapham, London SW4 6JX (GB).

(74) Agent: **ELRIFI, Ivor, R.**; Mintz, Levin, Cohn, Ferris, Glovsky, and Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: FAMILY-BASED ASSOCIATION TESTS FOR QUANTITATIVE TRAITS USING POOLED DNA

(57) Abstract: While SNP-based marker sets and population-level DNA repositories are approaching sufficient size for whole-genome association studies, individual genotyping remains very costly. Pooled DNA tests are a less costly alternative, but uncertainty about loss of power due to allele frequency measurement error and population stratification hinder their use. Here we describe how to optimize pooled tests as an explicit function of measurement error, and we present family-based tests that eliminate stratification effects. We show that identification of functional genetic variants and linked markers may be feasible with current-day instruments.

FAMILY-BASED ASSOCIATION TESTS FOR QUANTITATIVE TRAITS USING POOLED DNA

INTRODUCTION

Association tests of outbred populations are thought to have greater power than traditional family-based linkage analysis to identify the genetic variants contributing to complex human diseases (Risch and Merikangas, 1996; Ott 1999; Ardlie 2002). A genome scan based on allelic association would require approximately 100,000 markers, estimated by dividing the 3.3 gigabase human genome by the several kilobase extent of population-level linkage disequilibrium (Abecasis et al 2001; Reich et al. 2001). Single-nucleotide polymorphisms (SNPs) occur at sufficient density to provide a suitable marker set (Collins et al. 1997). Furthermore, SNPs in coding and regulatory regions have additional value as potential functional variants.

Individual genotyping remains prohibitively expensive for a genome scan. One method to reduce cost is to pool DNA from individuals with extreme phenotypic values and to measure the allele frequency difference between pools (Barcellos et al.,1997; Daniels et al., 1998; Fisher et al., 1999; Hill et al., 1999; Shaw et al., 1998; Stockton et al., 1998; Suzuki et al., 1998). Initial attention focused on pooled designs for dichotomous traits and case-control studies (Risch and Teng 1998). More recently, pooled tests have been discussed for quantitative traits, a more appropriate model for diseases such as obesity and hypertension. In the absence of experimental error, the optimal design for an unrelated population is to compare frequencies between pools of the most extreme 27% of individuals ranked by phenotypic value, retaining 80% of the information of individual genotyping (Bader et al., 2001). Experimental sources of error, primarily allele frequency measurement error, degrade the test power (Jawaid et al., 2002).

Population stratification poses a second challenge to practical use of pooled tests for human populations. Genomic control methods, developed to reduce stratification effects in genotype-based association tests (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Pritchard et al. 2001; Zhang and Zhou, 2001), are not directly applicable to pooled tests.

Here we present optimized pooled DNA test designs, including family-based tests robust to stratification. Estimates of test power explicitly include allele frequency measurement error. This distinguishes our treatment from prior theoretical work, permits the optimization of test design as a function of known parameters, and provides a bridge to experimentalists seeking practical guidance for whether to attempt and how to perform pooled association tests.

SUMMARY OF THE INVENTION

The invention is drawn to a method for detecting an association in a population of unrelated individuals between a genetic locus and a quantitative phenotype, wherein two or more alleles occur at the locus, and wherein the phenotype is expressed using a numerical phenotypic value whose range falls within a first numerical limit and a second numerical limit. This method comprises the steps of:

a) obtaining the phenotypic value for each individual in the population;

b) determining the minimum number of individuals from the population required for detecting the association using a non-centrality parameter;

c) selecting a first subpopulation of individuals having phenotypic values that are higher than a predetermined lower limit and pooling DNA from the individuals in the first subpopulation to provide an upper pool;

d) selecting a second subpopulation of individuals having phenotypic values that are lower than a predetermined upper limit and pooling DNA from the individuals in the second subpopulation to provide a lower pool;

e) for one or more genetic loci, measuring the frequency of occurrence of each allele at said locus in the upper pool and the lower pool;

f) for a particular genetic locus, measuring the difference in frequency of occurrence of a specified allele between the upper pool and the lower pool; and

g) determining that an association exists if the allele frequency difference between the pools is larger than a predetermined value.

In one embodiment of the invention, the difference in frequency of occurrence of the specified allele has associated with it an error of measurement. In one aspect of the invention the error of measurement is 0.04. In another, the error of measurement is 0.01.

In another embodiment of the invention, the predetermined lower limit is set so that the upper pool ranges from including the highest 37% of the population to including the highest 19% of the population and the predetermined upper limit is set so that the lower pool ranges from including the lowest 37% of the population to including the lowest 19% of the population. In another aspect of the invention, the predetermined lower limit is set so that the upper pool includes the highest 27% of the population and the predetermined upper limit is set so that the lower pool includes the lowest 27% of the population.

In another embodiment of the invention, the genetic locus has two alleles.

In another embodiment of the invention, the population includes individuals who may be classified into classes. In one aspect of the invention, the classes are based on an age group, gender, race or ethnic origin. In another aspect of the invention, the members of a class are included in the pools.

In another embodiment of the invention the method is used for determining the genetic basis of disease predisposition.

In another embodiment of the invention, the genetic locus which is analyzed for determining the genetic basis of disease predisposition contains a single nucleotide polymorphism.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. The information retained by the between-family pooled test design, expressed as a fraction of the information from individual genotyping followed by a between-family test, is depicted sibships of size 4, 2, and 1, each population having 1000 total individuals. The optimal pooling fraction, indicated by an arrow, shifts to lower values as the number of sibs per family decreases. The optimal fraction and corresponding information retained also shift to lower values as the minor allele frequency decreases, with results shown for frequencies 0.1 and 0.01. The raw measurement error is 0.01.

Figure 2. The optimal number of sibs to select from each family (top panel) and the information retained relative to individual genotyping (bottom panel) are shown for sibship sizes 2–5, 6, 8, 16, and 32 as a function of the scaled measurement error κ. For sibships through 5, it is always optimal to select just the highest and lowest sib.

Figure 3. The optimal fraction of families to select (top panel) and information retained (lower panel) are displayed for sibships of size 2 through 6 as a function of the scaled measurement error κ.

Figure 4. The optimal pooling fraction (top panel) and information retained (bottom panel) for between-family and within-family tests of a population of 500 sib-pairs are shown as a function of raw measurement error for marker frequencies 0.5 and 0.01. The within-family tests include pre-selection of discordant-like families.

Figure 5. The optimal pooling fraction (top panel) and the information retained (bottom panel) from exact numerical calculations (solid line) and an analytical fit (dashed line) are displayed as a function of the normalized measurement error κ. The fit coincides with the exact results for the information retained.

DETAILED DESCRIPTION

We present optimized designs for pooled DNA tests conducted on a population of N/s families, each a sibship of size s (N total individuals). The genotypic correlation within a sibship is denoted r, with typical values of 1/4, 1/2, and 1 for half-sibs, full-sibs, and monozygotic twins. Sibships may also represent inbred lines; in this case, r is the genetic correlation within each line. Sibs in different families are assumed to have uncorrelated genotypes.

To conduct a pooled DNA test for association of a particular allele A_1 with a quantitative trait, individuals are selected for an upper pool, comprising higher phenotypic values, and a lower pool, comprising lower phenotypic value, using designs reminiscent of selection strategies for optimizing breeding value and for QTL mapping (Hill 1971; Kimura and Crow 1978; Ollivier et al. 1997). We restrict attention to balanced designs in which each pool has fN individuals, with $f \leq 0.5$ defined as the pooling fraction. Balanced designs are favored when high and low phenotypes are treated symmetrically; asymmetry can favor unbalanced designs (Jawaid et al., 2002).

We consider four designs: (i) unrelated individuals ($s = 1$), in which the fN individuals having highest and lowest phenotypic values are selected for the upper and lower pools respectively; (ii) between-family, in which all s sibs from the fN/s families having highest and lowest mean phenotypic values are selected for the upper and lower pools; (iii) within-family, in which the s' sibs having highest and lowest phenotypic values within each family are selected for the upper and lower pools, yielding a pooling fraction $f = s'/s$; (iv) within-family with pre-selection of discordant families, in which a fraction f' of families with greatest within-family phenotypic variance are selected, $\mathrm{Var} = \sum_s \left(X_s - \overline{X}\right)^2$ where X_s is the phenotype of sib s and $\overline{X}$ is the family mean, then the extreme high and low sib within each selected family are selected for the upper and lower pool for a final pooling fraction $f = f'/N$.

A suitable statistic for a two-sided test for each design is

$$Z^2 = \frac{(\hat{p}_U - \hat{p}_L)^2}{\mathrm{Var}(\hat{p}_U - \hat{p}_L)},$$

where the estimated frequencies of allele A_1 in the upper and lower pools are denoted $\hat{p}_U$ and $\hat{p}_L$. The variance is the sum of three terms, $\mathrm{Var}(\hat{p}_U - \hat{p}_L) = V_S + V_C + V_M$. The sampling variance V_S represents the unavoidable error in estimating the population frequency from a finite sample. The concentration variance V_C arises from sample-to-sample DNA concentration variance within a pool. The measurement variance is $V_M = 2\varepsilon^2$, where ε is the experimental allele frequency measurement error for each pool. We assume that the three sources of variation are independent, which should be justified when individual and pooled DNA samples are treated uniformly. In an ideal experiment, V_C and V_M vanish, and the total variance is V_S.

Under the null hypothesis, Z^2 has a χ^2 distribution with one degree of freedom. Under the alternate hypothesis, the tested marker is assumed to be a bi-allelic quantitative trait locus (QTL) with alleles A_1 and A_2 occurring at frequencies p and $(1-p) \equiv q$. For between-family tests, the alleles are assumed to be in Hardy-Weinberg equilibrium and the population is assumed to be random mating; these assumptions may be relaxed for within-family tests. The variance of the allele frequency per individual is $\sigma_p^2 = pq/2$. For each design, the allele frequency is estimated as $\hat{p} = (\hat{p}_U + \hat{p}_L)/2$. The estimated variance of the allele frequency per individual is denoted $\hat{\sigma}_p^2$ and equals $\hat{p}(1-\hat{p})/2$.

The mean phenotypic effects are $m_G = a, d,$ and $-a$ for genotypes $G = A_1A_1$, A_1A_2, and A_2A_2, respectively. The dominance ratio d/a describes the inheritance mode with typical values -1, 0, and 1 for pure recessive, additive, or dominant inheritance. The proportion of trait variance accounted for by the QTL is denoted σ_Q^2,

$$\sigma_Q^2 = 2pq[a - d(p-q)]^2 + [2pqd]^2 = \sigma_A^2 + \sigma_D^2.$$

The mean QTL effect is $m = (p-q)a + 2pqd$. Phenotypic values are assumed to be normally distributed for each genotype with mean $\mu_G = m_G - m$ and residual variance $\sigma_R^2 = 1 - \sigma_Q^2$ arising from all genetic and environmental factors other than the QTL. The distribution of phenotypic values in the population is a mixture of three normal distributions with overall mean 0 and variance 1. The total phenotypic correlation between sibs from genetic factors (including the QTL) and environmental factors is termed t.

1

The non-centrality parameter (*NCP*),

$$NCP = [\mathrm{E}(\hat{p}_U - \hat{p}_L)]^2 / \mathrm{Var}(\hat{p}_U - \hat{p}_L) \ ,$$

measures the information provided by a pooled DNA test. The notation $\mathrm{E}(\hat{O})$ is the expectation of an observable $\hat{O}$. The approach followed below is to evaluate the numerator of the *NCP* as a function of the model parameters, providing accurate analytical results when possible and simulation results otherwise. For the denominator of the *NCP*, analytical results are obtained for the null hypothesis. For the alternative hypothesis, the expected allele frequencies for each pool have offsetting changes from p to $p \pm \delta p$ (see Methods for derivation), and the value of the denominator decreases by a small value proportional to $(\delta p/p)^2$. We make a conservative approximation by ignoring the change and using the null hypothesis denominator for the alternative hypothesis as well. In this case, the *NCP* equals $(z_{\alpha/2} - z_{1-\beta})^2$, where α and β are the type I and II error rates for the two-sided test. Maximizing the *NCP* optimizes the test.

The denominator of the *NCP* is shown in the Methods to have the form

$$V_S + V_C + V_M = \frac{2G\hat{\sigma}_p^2}{Nf} + \frac{2\tau^2\hat{\sigma}_p^2}{Nf} + 2\varepsilon^2 = \frac{2\hat{\sigma}_p^2}{Nf}\cdot\left(G+\tau^2\right)\cdot\left[1+\frac{Nf\varepsilon^2}{\left(G+\tau^2\right)\hat{\sigma}_p^2}\right]$$

$$= \frac{2\hat{\sigma}_p^2}{Nf}\cdot\left(G+\tau^2\right)\cdot\left(1+f\kappa^2\right)$$

where τ is the coefficient of variation for DNA concentration. The constant *G* depends only on the family structure and equals 1 for pools of unrelated individuals, *sR* for the between-family design, and (1–*r*) for both within-family designs; the standard notation *R* relates the sib genotypic correlation *r* to family-based variance components,

$$R = \frac{1}{s}\cdot\left[1+(s-1)r\right].$$

Typically τ is less than 10%; τ^2 may usually be ignored relative to *G*. The term κ^2 is used as shorthand,

$$\kappa^2 \equiv \varepsilon^2 / \left[\left(G+\tau^2\right)\hat{\sigma}_p^2 / N\right].$$

Referred to as the scaled measurement error, κ represents the raw measurement error, ε, scaled by the remaining sources of error in the allele frequency difference. In practice, κ can be calculated prior to pooling because it depends on known quantities.

The numerator of the *NCP* is shown in the Methods to have the form

$$\left[\mathrm{E}(\hat{p}_U - \hat{p}_L)\right]^2 = \frac{4\sigma_A^2\hat{\sigma}_p^2\phi\left[\Phi^{-1}(1-f)\right]^2}{\sigma_R^2 f^2}\cdot F$$

where $\phi(z)$ is the normal density $(2\pi)^{-1/2}\exp(-z^2/2)$, $\Phi(z)$ is the cumulative normal probability and $\Phi^{-1}(z)$ its functional inverse. The constant F equals 1 for pools of unrelated individuals, R^2/T for between-family pools, and $(1-R)^2/(1-T)$ for within-family pools without pre-selection. For the within-family design using discordant-like pre-selection, $F = (1-r)^2/2(1-t)$ for sib-pairs (expressions for larger sibships are unwieldy). The term R has the same definition as before, and T is the standard factor relating the sib phenotypic correlation t to family-based variance components, $T = \frac{1}{s}\cdot\left[1+(s-1)t\right]$.

Combining terms, the analytical result for the NCP, valid for small QTL effect, is

$$NCP = \frac{N\sigma_A^2}{\sigma_R^2}\cdot\frac{F}{G+\tau^2}\cdot\frac{2\phi\left[\Phi^{-1}(1-f)\right]^2}{f+f^2\kappa^2}.$$

The first factor is identical to the NCP for an association test performed by individual genotyping on a population of N unrelated individuals; the second factor, with $\tau = 0$, is the correction for individual genotyping a population of N/s families each having s sibs and then performing either a between-family test, with $F/G = R/sT$, or a within-family test, with $F/G = (s-1)R/s(1-T)$. The third factor represents the fraction of information retained when the association test is performed by pooling instead of individual genotyping, and maximizing this factor with respect to the pooling fraction f provides the optimal pool size. When the measurement error $\varepsilon = 0$, tests are optimized with $f = 0.27$ and 80% of the information is retained (Bader et al. 2001). As ε increases, the maximum information that can be retained is determined entirely by the single collective term κ.

Expressions for F, G, and κ^2 are summarized in Table I, and we now provide examples of each family-based design. Information retained by the between-family design is depicted in Fig. 1, with results for 3 sibship sizes: sib-quads, sib-pairs, and unrelated individuals, each population having 1000 total individuals. The optimal pooling fraction, indicated by an arrow, shifts to lower values as the number of sibs per family decreases. The optimal fraction and corresponding information retained also shift to lower values as the minor allele frequency decreases, with results shown for frequencies 0.1 and 0.01. The raw measurement error $\varepsilon = 0.01$ in this example, and the pooling fraction and information retained would decrease for larger ε (see Fig. 4 for examples of changing ε).

In Fig. 2, the optimal number of sibs to select from each family (top panel) and the information retained relative to individual genotyping (bottom panel) are shown as a function of the scaled measurement error κ for sibship sizes of 2–5, 6, 8, 16, and 32. For sibships through 5, it is always optimal to select just the highest and lowest sib. For larger families and small measurement error, the top and bottom quarters of the sibs are pooled and 80% of the information is retained. The pooling fraction and information decrease as the measurement error increase.

Within-family tests can be improved by pre-selection of discordant-like families, as shown in Fig. 3. The optimal fraction of families to select (top panel) and information retained (bottom panel) are displayed for sibships of size 2 through 6 as a function of the scaled measurement error κ (results determined by computer simulation). The fraction of families and information retained both decrease as κ increases. Discordant pre-selection has the greatest benefit for sib-pairs: for the smallest values of κ, only 56% of families are selected, retaining 80% of the information; had all families been used, only 60% of the information would have been retained. Pre-selection is less important for trios and larger sibships.

In Fig. 4, the optimal pooling fraction (top panel) and information retained (bottom panel) using between-family pools and using within-family pools with discordant-like pre-selection are displayed for a population of 500 sib-pairs (1000 individuals) as a function of the raw measurement error ε. Results are shown marker frequencies 0.5 and 0.01. With no measurement error, the optimal pooling fraction of 0.27 retains 80% of the information in each case. As measurement error increases, the optimal pooling fraction decreases, as does the information retained.

The information loss increases for rarer alleles and is worse for the within-family test than for the between-family test. This behavior can be deduced from the scaled error κ^2, which is inversely proportional to the allele frequency sampling variance. Since the sampling variance is 3× smaller within-family vs. between-family, κ^2 is 3× larger, $4N\varepsilon^2/p(1-p)$ vs. $4N\varepsilon^2/3p(1-p)$, and more information is lost. The inverse dependence of κ^2 on the allele frequency explains the decrease in power for rare alleles.

Because the allele frequency difference between sibs is uncorrelated from their allele frequency mean, the between-family and within-family tests are independent estimators of σ_A even when individuals contribute their DNA under both designs. The *NCP* of a combined test is the sum of the *NCP*s for each test and it too follows a χ^2 distribution with 1 degree of freedom. In practice, estimates for σ_A may obtained by inverting the expressions for $\mathrm{E}(\hat{p}_U - \hat{p}_L)$ provided in Table I, then weighting each estimator by the inverse of its variance.

Population stratification may be indicated by a difference between the estimates for σ_A from a between-family and within-family test. In the absence of stratification, the difference follows a normal distribution with variance

$$\mathrm{Var}\left[\hat{\sigma}_{A+} - \hat{\sigma}_{A-}\right] = V_+ \cdot \left[f_+^2 T \sigma_R^2 / 4y_+^2 R^2 \hat{\sigma}_p^2\right] + V_- \cdot \left[f_-^2 (1-T)\sigma_R^2 / 4y_-^2 (1-R)^2 \hat{\sigma}_p^2\right]$$

where the "+" and "–" subscripts refers to the between-family and within-family designs respectively, $y_\pm = \phi\left[\Phi^{-1}(1-f_\pm)\right]$, and V represents the total variance, $V_S + V_C + V_M$, for each design. When stratification is indicated, the between-family estimate of σ_A may be unreliable but the within-family estimate remains robust.

In Fig. 5, the optimal pooling fraction (top panel) and the information retained (bottom panel) are displayed as a function of the scaled measurement error κ. The information retained is calculated assuming no concentration variance. In addition to the numerically calculated results, an accurate fit is shown using the functional form

$$f = 1 - \Phi\left[A - (3/A)\ln A - 0.067\right], \text{ with}$$

$$A(\kappa) = \sqrt{2 + \ln\left(1 + 3\kappa^2 + \frac{2}{\pi}\kappa^4\right)}.$$

A justification for this functional form is provided in the Methods. The greatest deviations for the pooling fraction are at κ = 0.5, where the fit yields a pooling fraction that is 0.006 too high, and at κ = 3.5, where the fit is 0.01 too low. The information retained using the analytical value for the pooling fraction coincides with the exact numerical results on the scale of the figure. The experimental measurement error ε corresponding to the scaled error κ depends on the population structure and marker frequency. For example, for a population of 500 cases, 500 matched unrelated controls, and 10% marker frequency, ε = 0.0067κ is the raw error corresponding to κ.

Based on the pooled designs described above, we outline a prospective study using 100,000 markers to detect QTLs with a 1% effect. If 100 false-positives are permitted from pooled tests (the false-positives may be resolved using individual genotyping) and 80% power is required, the *NCP* is 17. We assume pooling of discordant sib-pairs to protect against stratification effects. At the scaled error κ = 1 where the pooled tests are still close to maximum power, the pooling fraction would be 21%, 65% of the information of a population would be retained, and a population of 2600 individuals would be required. The raw measurement error corresponding to $\kappa = 1$ for this population size is 0.005 for an allele with 50% frequency and 0.002 for an allele with 5% frequency, $5\times$ to $10\times$ more precise than achieved by current-day instrumentation.

We can account for current-day precision by setting $\kappa = 10$, which from Fig. 5 is seen to retain 7.7% of the information and corresponds to a pooling fraction of 1.6% of a total population of 22,000. In this case, the precision required for a pooled test is 0.017 for an allele with 50% frequency and 0.007 for an allele with 5% frequency. These precisions are within the range of current performance, especially if repeated measures are used to decrease the effective measurement error. The cost to collect and score such a population for multiple disease-related phenotypes would be under $50 million. Selection schemes could then be applied to generate pools for each phenotype in turn.

As noted previously, pooled tests perform worse for within-family tests and rare alleles, and may therefore be difficult to apply to disease-risk variants under negative selection pressure. The loss of power may be less severe for pharmacogenetic studies of variants affecting drug response, where selection pressure is absent, and for test crosses of model organisms (Grupe et al. 2001) or agricultural species whose marker frequencies are under experimental control.

The analysis provided here for quantitative traits may be extended to threshold characters yielding dichotomous classifications of a population. For case-control classification, the disease prevalence corresponds to the pooling fraction f. When the quantitative character is available for measurement, it is approximately 4× more efficient to compare unrelated individuals with extremely high vs. extremely low characters than to compare the derived cases vs. controls (Bader et al. 2001).

In summary, we have derived the optimal pooling fractions for within-family and between-family tests of association. With ideal instrumentation, 80% of the information is retained and the optimal pooling fraction is 27%. As allele frequency measurement error increases, the optimal pooling fraction and the information retained both decreases. The information loss is more severe for low-frequency alleles and for within-family tests. The optimal pooling fraction depends on a single parameter representing the measurement error, and optimized pooling designs are provided as a function of this parameter.

EXAMPLES

Example 1: Sampling variance and concentration variance

Let p_i represent the frequency of allele A_1 for individual i, either 0, 1/2, or 1, and c_i represent the concentration of DNA contributed by this individual to a pool of n individuals. Neglecting measurement error, the allele frequency p^* for the pool is

$$\begin{aligned} p^* &= \sum \frac{c_i p_i}{\sum c_j} = p + \sum \frac{(c_0 + \delta c_i)\delta p_i}{\sum c_0 + \sum \delta c_j} \\ &= p + \sum \frac{\left(\frac{1}{n} + \frac{\delta c_i}{n c_0}\right)\delta p_i}{1 + \frac{1}{n c_0}\sum \delta c_j} \\ &\approx p + \sum \delta p_i \left(\frac{1}{n} + \frac{1}{n}\frac{\delta c_i}{c_0}\right)\left(1 - \frac{1}{n}\sum \frac{\delta c_j}{c_0}\right) \\ &\approx p + \frac{1}{n}\sum \delta p_i + \sum \delta p_i \left(\frac{\delta c_i}{n c_0} - \sum \frac{\delta c_j}{n^2 c_0}\right) \\ &\equiv p + \frac{1}{n}\sum \delta p_i + \frac{1}{n}\sum \delta p_i \delta c_i' \end{aligned}$$

which defines the relative concentration error $\delta c_i'$. The terms δp_i and $\delta c_i'$ are uncorrelated, and each has expectation zero. Furthermore, the sum of the $\delta c_i'$ terms is constrained to be zero. The variance of p^* is

$$\begin{aligned} \mathrm{Var}(p^*) &= \frac{1}{n^2}\sum_{i,j} \mathrm{Cov}(\delta p_i, \delta p_j) + \frac{1}{n^2}\sum_{i,j} \mathrm{Cov}(\delta c_i', \delta c_j')\mathrm{Cov}(\delta p_i, \delta p_j) \\ &= \frac{1}{n^2}\sum_{i,j} r_{ij}\sigma_p^2 + \frac{\tau^2}{n}\sigma_p^2 \end{aligned}$$

We have used

$$\mathrm{Cov}(\delta p_i, \delta p_j) = \frac{p(1-p)}{2} r_{ij} = \sigma_p^2 r_{ij} \text{ and}$$

$$\mathrm{Cov}(\delta c_i', \delta c_j') = \tau^2\left(\delta_{ij} - \frac{1}{n}\right) \approx \tau^2 \delta_{ij},$$

with the concentration coefficient of variation defined as $\tau \equiv [\mathrm{Var}(c_i)]^{1/2}/c_0$ and the genotypic correlation between a pair of individuals defined as r_{ij}.

For the between-family design, a pool of n individuals contains n/s sibships of size s and genotypic correlation r. The result for $\mathrm{Var}(p^*)$ is

$$\mathrm{Var}(p^*) = \frac{sR}{n}\sigma_p^2 + \frac{\tau^2}{n}\sigma_p^2,$$

with $R = (1/s)[1+(s-1)r]$. Since the individuals in the upper and lower pools are unrelated, $V_S + V_C = 2\,\mathrm{Var}(p^*)$.

For a within-family design, the allele frequency difference between pools is

$$\Delta p^* = \frac{1}{n}\sum_i (1+\delta c_i')\delta p_i - \frac{1}{n}\sum_j (1+\delta c_j')\delta p_j,$$

where i and j label individuals in the upper and lower pools respectively. The variance is

$$\begin{aligned}\mathrm{Var}(\Delta p^*) &= \frac{2}{n^2}\sum_{i,i'}\mathrm{Cov}(\delta p_i, \delta p_{i'})[1+\mathrm{Cov}(\delta c_i', \delta c_{i'}')] - \frac{2}{n^2}\sum_{i,j}\mathrm{Cov}(\delta p_i, \delta p_j)\\ &= \frac{2(1-r)}{n}\sigma_p^2 + \frac{2\tau^2}{n}\sigma_p^2.\end{aligned}$$

Example 2: Expected allele frequency difference and non-centrality parameter

The genotype-dependent phenotype distribution is defined using a variance components model,

$$X_{ki} = Y_k + Y_{ki} + \mu_{ki}.$$

Family and individual effects are normally distributed with mean zero and variance

$$\begin{aligned}\mathrm{Var}(Y_k) &= t - r\sigma_A^2 - u\sigma_D^2\\ \mathrm{Var}(Y_{ki}) &= \sigma_R^2 - t + r\sigma_A^2 + u\sigma_D^2\end{aligned}$$

The family index is k, the sib index is i, and the individual phenotypes X_{ki} are the sum of Y_k, the family effect excluding the QTL, Y_{ki}, the individual effect excluding the QTL, and μ_{ki}, the QTL effect $\mu(G_{ki})$ for sib i. The total phenotypic correlation between sibs is t. Both r and u relate to the genetic background shared between sibs, r being the genotypic correlation (1 for monozygotic twins, 1/2 for full sibs, 1/4 for half sibs) and u being the shared genotype expectation (1 for monozygotic twins, 1/4 for full sibs, 0 for half sibs) (Falconer and Mackay 1996).

The observed phenotypes X_{ki} are re-expressed as family means and individual deviations from family means,

$$X_{k\bullet} = \frac{1}{s}\sum_i X_{ki}$$
$$\delta X_{ki} = X_{ki} - X_{k\bullet}.$$

Similar quantities are defined for the QTL effects,

$$\mu_{k\bullet} = \frac{1}{s}\sum_i \mu_{ki}$$
$$\delta\mu_{ki} = \mu_{ki} - \mu_{k\bullet},$$

and the variances of the observed quantities excluding QTL effects are

$$\mathrm{Var}(X_{k\bullet} - \mu_{k\bullet}) = \frac{1}{s}\left[\sigma_R^2 + (s-1)(t - r\sigma_A^2 - u\sigma_D^2)\right] \equiv T\sigma_R^2$$
$$\mathrm{Var}(\delta X_{ki} - \delta\mu_{ki}) = (1-T)\sigma_R^2$$

When the QTL effects are small, $T \approx (1/s)[1 + (s-1)t]$ is an accurate approximation.

The probability that sibling 1 from family *k* with genotypes $\mathbf{G} = (G_1, G_2, \ldots, G_s)$ is selected for the upper pool is $1 - \Phi[(X' - \mu_G)/\sigma]$, where $\Phi(z)$ is the cumulative normal probability. The variable under selection, denoted *X*, is either $X_{k\bullet}$ (between-family pools) or δX_{k1} (within-family pools); μ_G is either $\mu_{k\bullet}$ (between-family pools) or $\delta\mu_{k1}$ (within-family pools); the variance of $X - \mu_G$ is σ^2, either $T\sigma_R^2$ (between-family pools) or $(1-T)\sigma_R^2$ (within-family pools) ; and *X'* is the selection threshold applied to *X*. Because the labeling of sibs is arbitrary, the fraction *f* of individuals selected for pooling is equal to the probability that sib 1 is selected, i.e. the probability that *X* is greater than the selection threshold,

$$f = \sum_{\mathbf{G}} \Pr(\mathbf{G})\{1 - \Phi[(X' - \mu_G)/\sigma]\},$$

where Pr(**G**) is the probability of observing the sibship genotypes **G**.

To calculate the allele frequency of the selected individuals, the threshold *X'* is required as a function of *f*. Numerical inversion may be applied to the above equation. Alternatively, when the QTL effect is small ($\mu_G < \sigma$), the linear approximation

$$\Phi[(X' - \mu_G)/\sigma] \approx \Phi(X'/\sigma) - (\mu_G/\sigma)\phi(X'/\sigma)$$

is accurate, where $\phi(z) = d\Phi(z)/dz$ is the normal probability density. The terms linear in μ_G cancel in the sum over G, yielding $f = 1 - \Phi(X'/\sigma)$.

The expected allele frequency of the resulting pool is

$$E(\hat{p}_U) = \frac{1}{f}\sum_{G}\Pr(G)p_G \cdot \{1 - \Phi[(X' - \mu_G)/\sigma]\},$$

where p_G represents the allele frequency of sib 1. Using the linear expansion for $\Phi[(X' - \mu_G)/\sigma]$ yields

$$E(\hat{p}_U) = \sum_{G}\Pr(G)p_G + \frac{\phi(X'/\sigma)}{f\sigma}\sum_{G}\Pr(G)p_G\mu_G = p + \frac{\phi(X'/\sigma)}{f\sigma}E(p_G\mu_G).$$

An analogous expression for the lower pools gives a symmetric result, yielding

$$E(\hat{p}_U - \hat{p}_L) = \frac{2\phi[\Phi^{-1}(1-f)]}{f\sigma}E(p_G\mu_G)$$

where X'/σ has been replaced by $\Phi^{-1}(1-f)$.

The expectation of the correlation between p and μ for an individual is

$$\begin{aligned} E(p\mu) &= p^2[a-(p-q)a-2pqd] + 2pq\cdot\frac{1}{2}\cdot[d-(p-q)a-2pqd] \\ &= pq[a-(p-q)d] \\ &= \sigma_p\sigma_A \end{aligned}$$

Similarly, the correlation between sibs i and j is $E(p_i\mu_j) = r_{ij}\sigma_p\sigma_A$, where r_{ij} is their genotypic correlation. Summing over sibs yields either $R\sigma_p\sigma_A$ (between-family pools) or $(1-R)\sigma_p\sigma_A$ (within-family pools) for $E(p_G\mu_G)$, with $R = (1/s)[1+(s-1)r]$ as before.

Selecting discordant-like sib-pairs is equivalent to selection based on $|\delta X_{ki}|$, and the within-family analytical results are directly applicable. For larger families, discordant-like families are pre-selected in decreasing rank order of the within-family phenotypic variance $\sum_s \delta X_{ks}^2$ summed over siblings s.

We have ascertained that the analytical results for the *NCP* are virtually indistinguishable from exact numerical results when the QTL effect is 5% or less of the trait variance. For larger effects, roughly when the effect size σ_A^2 approaches the minor allele frequency, the genotype-dependent phenotype distributions become resolved, transforming a complex trait into Mendelian trait amenable to traditional linkage analysis.

Example 3: Analytical fit for the optimal pooling fraction

Optimizing the pooling fraction is equivalent to maximizing the objective function $I = 2y^2/(f + f^2\kappa^2)$, where y is shorthand for $\phi[\Phi^{-1}(1-f)]$. Writing f as $1-\Phi(z)$ and optimizing using $dI/dz = 0$ yields

$$y \cdot (1 + 2f\kappa^2) - 2zf \cdot (1 + f\kappa^2) = 0.$$

We have used $y = \phi(z)$, $dy/dz = -yz$, and $df/dz = -y$.

When κ^2 is large, z is also large, and f may be replaced by its asymptotic expansion for large z, $f = y \cdot (z^{-1} - z^{-3})$. With this substitution, the optimum satisfies

$$\frac{z^3}{2y\kappa^2} = 1.$$

Taking the natural logarithm of both sides and equating exponents,

$$\frac{z^2}{2} + 3\ln z - \ln\left(\kappa^2\sqrt{2/\pi}\right) \equiv J(z) = 0.$$

When κ and z are both large, the term $3\ln z$ is asymptotically small, giving

$$z \sim \sqrt{\ln(2\kappa^4/\pi)} \equiv B(\kappa).$$

An improved fit is obtained by perturbation theory by writing

$$z = B(\kappa)[1 + b(\kappa)],$$

where $\lim_{\kappa \to \infty} b(\kappa) = 0$. Substituting this expression for z into $J(z)$ and simplifying,

$$B^2 b + 3\ln[B(1+b)] = 0,$$

which gives the asymptotic form $b = (3/B^2)\ln B$, or

$$z \sim B - (3/B)\ln B.$$

For clarity, the functional dependence of B and b on κ has been suppressed.

The asymptotic form provides a good fit when κ is much larger than 1 but not for smaller values. Since the asymptotic behavior for large κ is not affected by introducing terms of lower order in κ, the fit can be improved for small κ without degrading the fit at large κ by writing

$$z = A - (3/A)\ln A + a_1 \text{, where}$$

$$A(\kappa) = \sqrt{a_2 + \ln\left(1 + a_3\kappa^2 + \frac{2}{\pi}\kappa^4\right)}.$$

The constants a_1, a_2, and a_3 are then selected to fit the exact numerical results at particular values of κ. Fitting the results $z = 0.612$ at $\kappa = 0$ and $z = 0.8047$ at $\kappa = 1$ provides the particular parameters

$$a_1 = -0.067,\ a_2 = 2,\ a_3 = 3.$$

Table I. The non-centrality parameter for family-based pooled DNA designs [a]

Design	F	G
Unrelated individuals	1	1
Between-family	R^2/T	sR
Within-family	$(1-R)^2/(1-T)$	$1-r$
Within-family, discordant pre-selection [b]	$(1-r)^2/2(1-t)$	$1-r$

[a] The non-centrality parameter (*NCP*) is $[\mathrm{E}(\hat{p}_U - \hat{p}_L)]^2/\mathrm{Var}(\hat{p}_U - \hat{p}_L)$. The numerator is $F \cdot \left(4\sigma_A^2\sigma_p^2\phi[\Phi^{-1}(1-f)]^2/\sigma_R^2 f^2\right)$, where *F* is provided for each design, *f* is the pooling fraction, σ_A^2 and σ_R^2 are the additive and residual variance for a QTL with allele frequency *p*, σ_p^2 is $p(1-p)/2$, $\phi(z)$ is the normal probability density and $\Phi(z)$ is the cumulative normal probability. The denominator of the *NCP* is $[2(G+\tau^2)\sigma_p^2/Nf] + 2\varepsilon^2$, where *G* is provided for each design, τ is the coefficient of variation for DNA sample concentrations in the pool, *N* is the total number of individuals before selection, and ε is the raw measurement error. The combined expression for the *NCP* is $(N\sigma_A^2/\sigma_R^2)\cdot[F/(G+\tau^2)]\cdot\{2\phi[\Phi^{-1}(1-f)]^2/(f + f^2\kappa^2)\}$, where κ^2 is $N\varepsilon^2/[(G+\tau^2)\sigma_p^2]$ and κ is termed the scaled error. Each sibship has *s* sibs with genotypic correlation *r* and phenotypic correlation *t*; *R* and *T* are $(1/s)[1+(s-1)r]$ and $(1/s)[1+(s-1)t]$, respectively.

[b] Analytical results are for sib-pairs only. For larger families see numerical results (Fig. 3).

References

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya A, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson OC (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Gen 68:191-197

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3: 299-309

Bader JS, Bansal A, and Sham P (2001) Efficient SNP-based tests of association for quantitative phonotypes using pooled DNA. Genescreen (in press)

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Gen 61:734-747

Collins FS, Guyer MS, Chakarvarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 274:1580-1581

Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen MJ (1998) A simple method for analysing microsatellite allele image patterns generated from DNA pools and its applications to allelic association studies. American Journal of Human Genetics 62:1189-97

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:788-808

Falconer (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann Hum Gen 51: 227-33

Falconer DS, MacKay TFC (1996) Introduction to quantitative genetics. Boston: Addison-Wesley

Fisher PJ, Turic D, Williams NM, McGuffin P, Asherson P, Ball D, Craig I, Eley T, Hill L, Chorney K, Chorney MJ, Benbow CP, Lubinski D, Plomin R, Owen MJ (1999) DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. Hum Mol Gen 8: 915-22

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G (2001) In silico mapping of complex disease-related traits in mice. Science 292: 1915-1918

Hill, WG. (1971) Design and efficiency of selection experiments for estimating genetic parameters. Biometrics 27: 293-311

Hill L, Craig IW, Asherson P, Ball D, Eley T, Ninomiya T, Fisher PJ, Turic D, McGuffin P, Owen MJ, Chorney K, Chorney MJ, Benbow CP, Lubinski D, Thompson LA, Plomin R (1999) DNA pooling and dense marker maps: a systematic search for genes for cognitive ability. Neuroreport 10: 843-848

Jawaid A, Bader JS, Purcell S, Cherny SS, Sham P (2002) Optimal selection strategies for QTL mapping using pooled DNA samples. European Journal of Human Genetics (in press)

Kimura M, Crow JF (1978) Effect of overall phenotypic selection on genetic change at individual loci. Proc Natl Acad Sci USA 75: 6168-6171

Ollivier L, Messer LA, Rothschild MF, Legault C (1997) The use of selection experiments for detecting quantitative trait loci. Genet Res, Camb 69: 227-232

Ott J (1999) Analysis of Human Genetic Linkage. Third edition. Johns Hopkins University Press, Baltimore

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Gen 65: 220-228

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204

Risch N and Teng J (1998) The relative power of family-based and case-control designs for linkage diequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8:1273

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Satten GA, Flanders DW, and Yang Q (2001) Accounting for unmeasured population substructure inb case-control studies of genetic association using a novel latent-class model. Am J Hum Gen 68: 466-477

Schork NJ, Nath SK, Fallin D, Chakarvati A (2000) Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control Subjects. Am J Hum Gen 67: 1208-1218

Sham PC, SS Cherny, S Purcell, and JK Hewitt (2000) Power of linkage versus association analyses of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Gen 66: 1616-1630

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res 8: 111-123

Stockton DW, Lewis RA, Abboud EB, Al Rajhi A, Jabak M, Anderson KL, Lupski JR (1998) A novel locus for Leber congenital amaurosis on chromosome 14q24. Human Genetics 103: 328-333

Suzuki K, Bustos T, Spritz RA (1998) Linkage disequilibrium mapping of the gene for Margarita Island ectodermal dysplasia (ED4) to 11q23. American Journal of Human Genetics 63:1102-1107

Zhang S, Zhao H (2001) Quantitative similarity-based association tests using population samples. American Journal of Human Genetics 69: 601-614

We claim:

1. A method for detecting an association in a population of unrelated individuals between a genetic locus and a quantitative phenotype, wherein two or more alleles occur at the locus, and wherein the phenotype is expressed using a numerical phenotypic value whose range falls within a first numerical limit and a second numerical limit, the method comprising the steps of

a) obtaining the phenotypic value for each individual in the population;

b) determining the minimum number of individuals from the population required for detecting the association using a non-centrality parameter;

c) selecting a first subpopulation of individuals having phenotypic values that are higher than a predetermined lower limit and pooling DNA from the individuals in the first subpopulation to provide an upper pool;

d) selecting a second subpopulation of individuals having phenotypic values that are lower than a predetermined upper limit and pooling DNA from the individuals in the second subpopulation to provide a lower pool;

e) for one or more genetic loci, measuring the frequency of occurrence of each allele at said locus in the upper pool and the lower pool;

f) for a particular genetic locus, measuring the difference in frequency of occurrence of a specified allele between the upper pool and the lower pool; and

g) determining that an association exists if the allele frequency difference between the pools is larger than a predetermined value.

2. The method of claim 1, wherein the difference in frequency of occurrence of the specified allele has associated with it an error of measurement.

3. The method of claim 2, wherein the error of measurement is 0.04.

4. The method of claim 2, wherein the error of measurement is 0.01.

5. The method described in claim 1, wherein the predetermined lower limit is set so that the upper pool ranges from including the highest 37% of the population to including the highest 19% of the population and the predetermined upper limit is set so that the lower pool

ranges from including the lowest 37% of the population to including the lowest 19% of the population.

6. The method of claim 1, wherein the predetermined lower limit is set so that the upper pool includes the highest 27% of the population and the predetermined upper limit is set so that the lower pool includes the lowest 27% of the population.

7. The method of claim 1, wherein the genetic locus has two alleles.

8. The method of claim 1 wherein the population includes individuals who may be classified into classes.

9. The method of claim 8, wherein the classes are based on an age group, gender, race or ethnic origin.

10. The method of claim 8, wherein all the members of a class are included in the pools.

11. The method of claim 1 for determining the genetic basis of disease predisposition.

12. The method of claim 11, wherein the genetic locus which is analyzed for determining the genetic basis of disease predisposition contains a single nucleotide polymorphism.

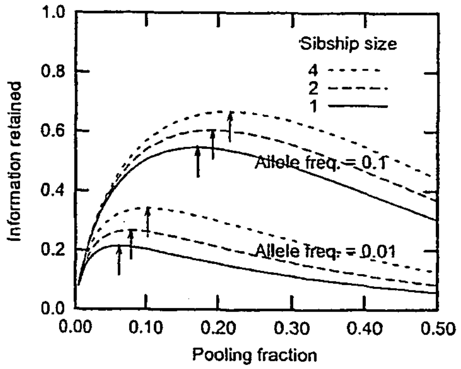
1.0
0.8
0.6
0.4
0.2
0.0
Information retained
0.00
0.10
0.20
0.30
0.40
0.50
Pooling fraction
Sibship size
4
2
1
Allele freq. = 0.1
Allele freq. = 0.01

Figure 1

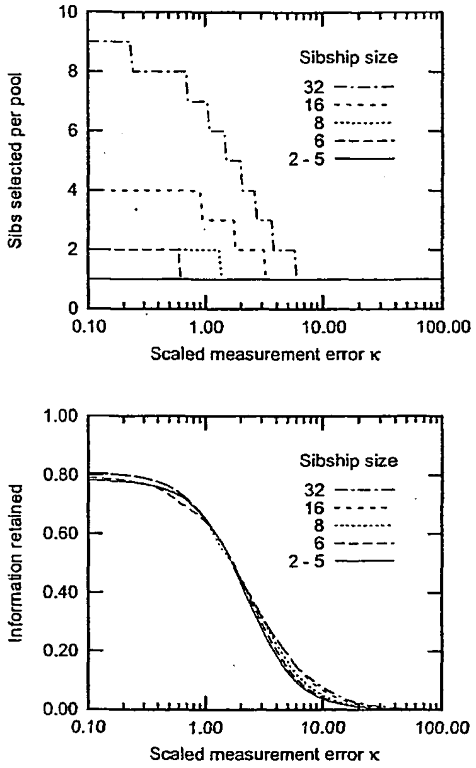
Sibs selected per pool
10
8
6
4
2
0
Sibship size
32
16
8
6
2 - 5
0.10
1.00
10.00
100.00
Scaled measurement error κ
Information retained
1.00
0.80
0.60
0.40
0.20
0.00
Sibship size
32
16
8
6
2 - 5
0.10
1.00
10.00
100.00
Scaled measurement error κ

Figure 2

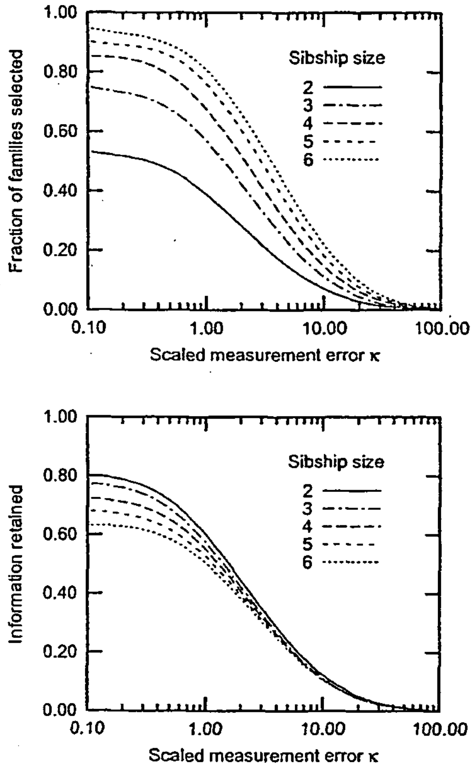
1.00
0.80
0.60
0.40
0.20
0.00
Fraction of families selected
Sibship size
2
3
4
5
6
0.10
1.00
10.00
100.00
Scaled measurement error κ
1.00
0.80
0.60
0.40
0.20
0.00
Information retained
Sibship size
2
3
4
5
6
0.10
1.00
10.00
100.00
Scaled measurement error κ

Figure 3

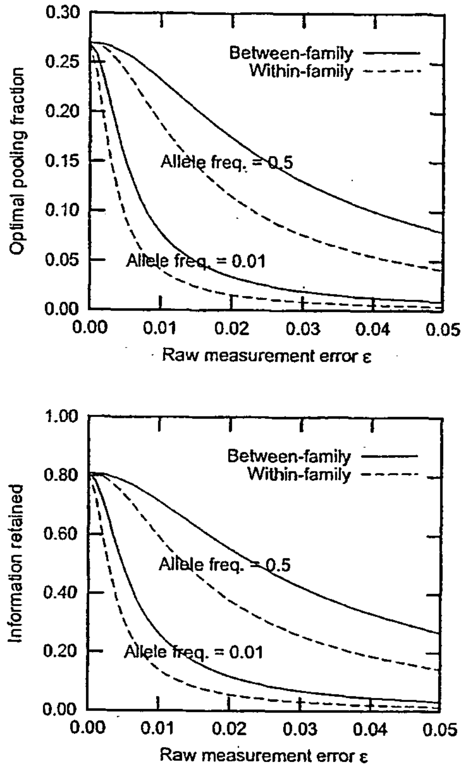
Optimal pooling fraction
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Between-family
Within-family
Allele freq. = 0.5
Allele freq. = 0.01
0.00
0.01
0.02
0.03
0.04
0.05
Raw measurement error ε
Information retained
1.00
0.80
0.60
0.40
0.20
0.00
Between-family
Within-family
Allele freq. = 0.5
Allele freq. = 0.01
0.00
0.01
0.02
0.03
0.04
0.05
Raw measurement error ε

Figure 4

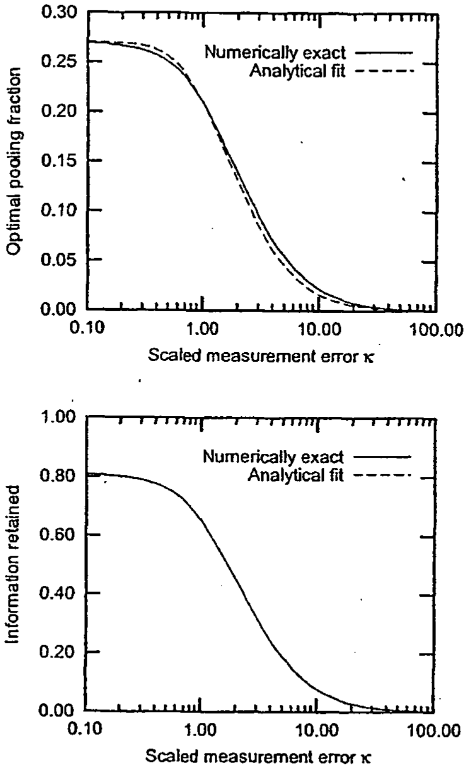
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Optimal pooling fraction
Numerically exact
Analytical fit
0.10
1.00
10.00
100.00
Scaled measurement error κ
1.00
0.80
0.60
0.40
0.20
0.00
Information retained
Numerically exact
Analytical fit
0.10
1.00
10.00
100.00
Scaled measurement error κ

Figure 5